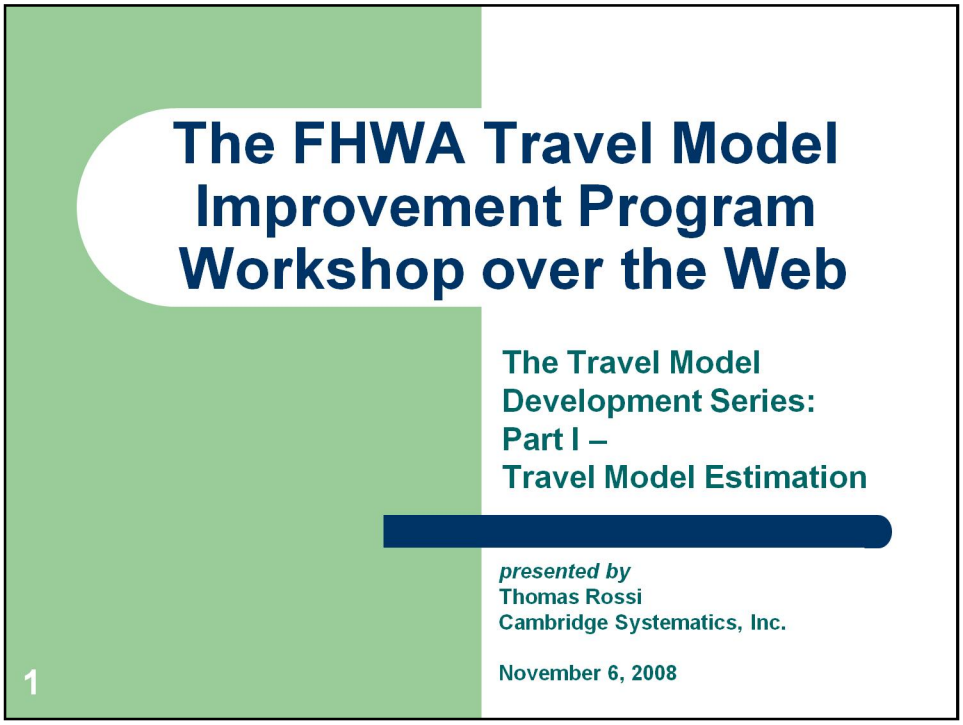
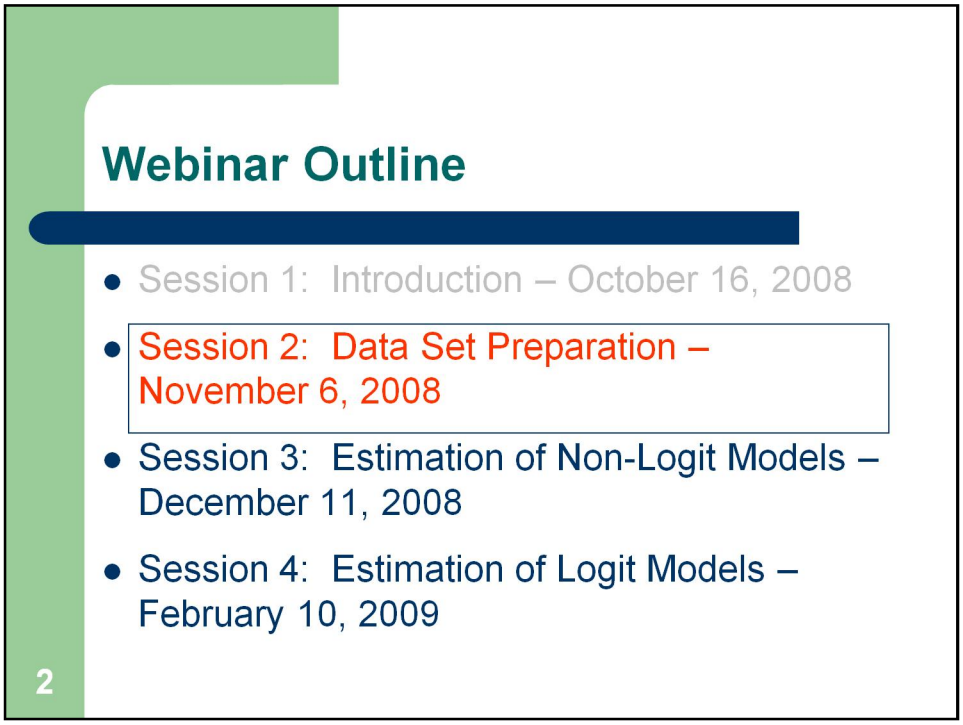


Key Message: Purpose of the Webinar Series

Details:

Welcome to the FHWA TMIP Workshop over the Web. This workshop is targeted at Transportation modelers who have a low to moderate level of familiarity with the estimation and validation of travel models.

This series of webinars will introduce the development of model estimation data sets, the structures of the various model components, and the procedures for estimating models. The workshop will include lectures, discussion, and "homework," that participants will be expected to complete between sessions.

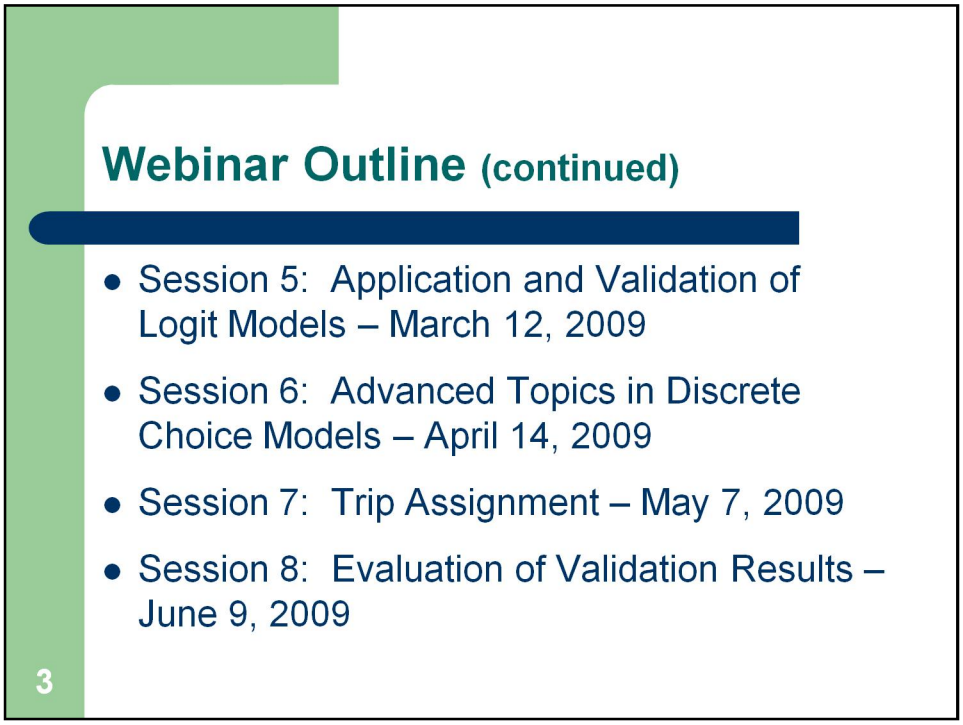


Key Message: Current Session

Details:

This session will deal with the nuances of data set preparation. Specifically, it will focus on the most common sources of data, the common pitfalls in survey data and some solutions to overcome these pitfalls.

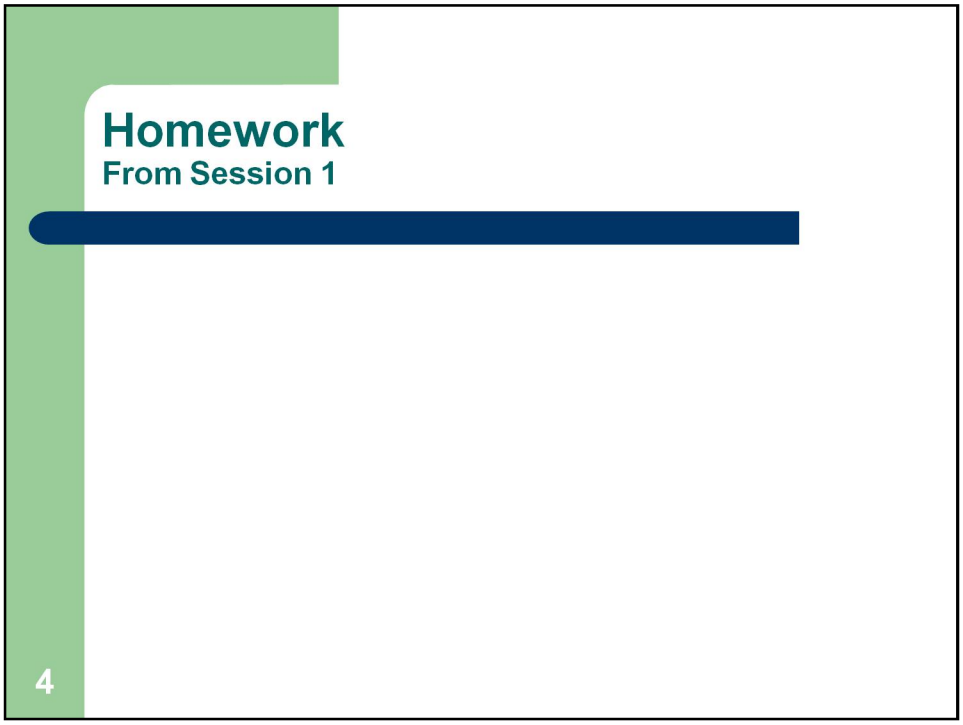
Session 3, which will be conducted on December 11, will cover the various aspects of non-logit model estimation. Session 4, will be conducted on February 2009 and will deal with logit estimation issues



Key Message: Upcoming Sessions

Details:

The dates for Sessions 5-8 have been determined. Session 5 will be conducted on March 12, 2009; Session 6 will be conducted on April 14, Session 7 on May 7, and Session 8 on June 9.

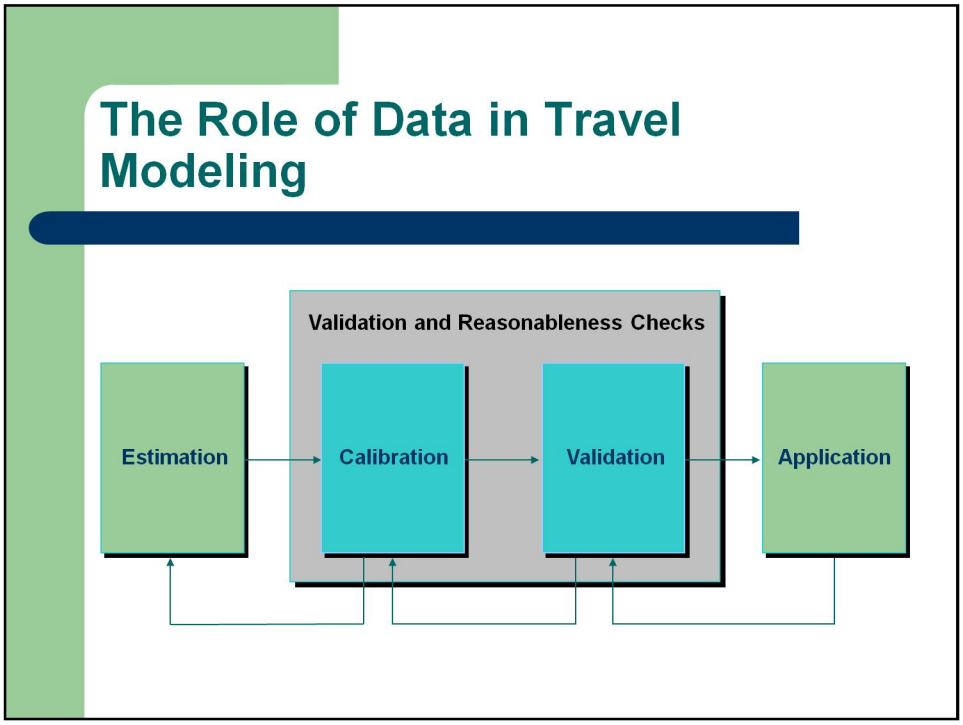


Key Message: Homework 1 Discussion

Details:

Please refer to the homework solutions posted at the following website.

http://tmip.fhwa.dot.gov/discussions/webinars/archive/tmw/downloads/homework_1.pdf



Key Message: The Various Stages in Modeling

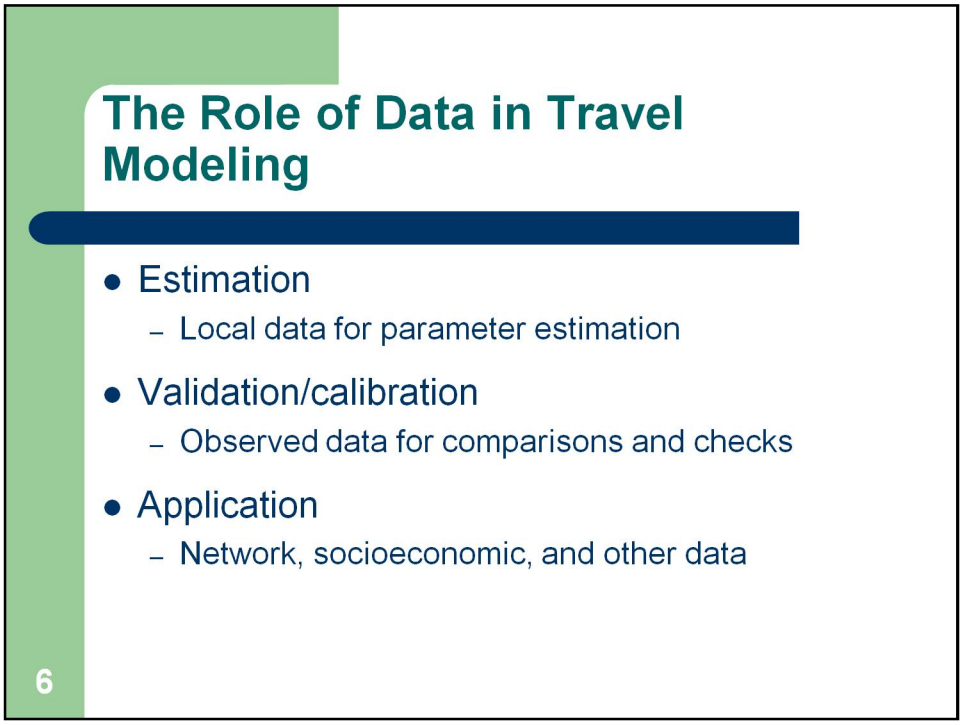
Details:

In model estimation, the parameters are developed. Commonly, parameters are estimated using statistical techniques using data from surveys or other sources.

The validation of the model involves checks of the parameters and the results, compared to other data or results from other model runs. Any issues identified during validation may result in adjustments to parameters (calibration). When parameters have been changed, the model must be revalidated.

The results of validation could involve re-estimation of parameters in some cases. For example, if the validation process identified that, say, mode choice results were correct overall but not by auto ownership level, the model might be reestimated with a new or revised variable dealing with auto ownership.

Model application can occur once the model has been validated for the base year. However, application for the forecast year is a necessary part of model validation and can lead to further calibration changes.

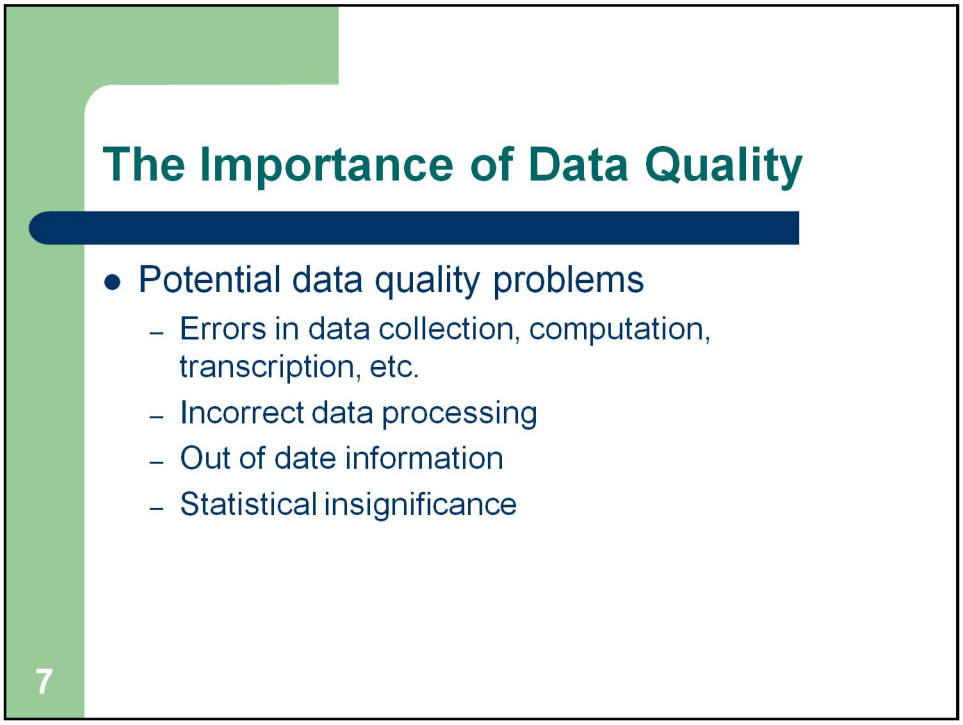


Key Message: Importance of Data

Details:

Good data are essential to the development of any travel model. The uses of data extend beyond model estimation; data are used equally extensively for validation and calibration purposes as well as for model application purpose.

Local survey data are typically used for the model parameter estimation. The local survey data can also be used as a reality check for the model results. Other data such as traffic counts are also often needed to check the performance of a model. Population and Employment data for base and future years are also essential to apply the model. Similarly, base and future year networks are essential to assign trip tables and generate link-wise traffic volume estimates.



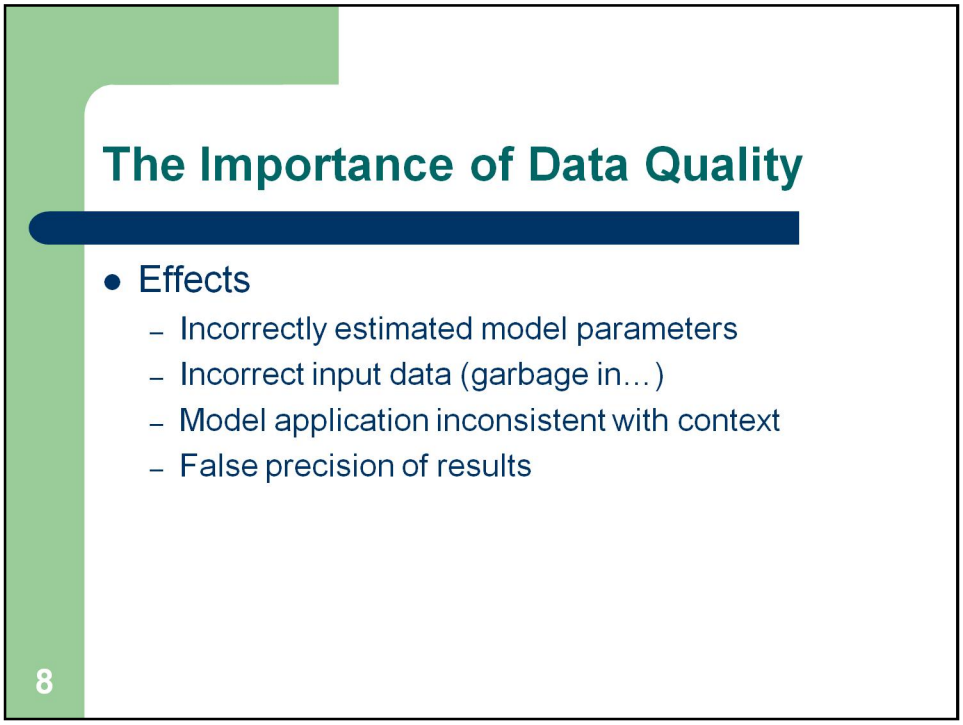
Key Message: Data Quality

Details:

Errors are virtually unavoidable in any data source. The data errors could arise during the data collection process itself through incorrect transcription of reported data, or through incorrect imputation of correctly transcribed data.

Even if the transcription and imputation are both right, data could be outdated and may not capture current travel behavior.

Another major problem with data is that it may not be representative of the true universe. This could happen in one or more ways. First, the data sample may not be large enough, thereby providing statistically insignificant results. Second the sampling plan may not be reflective of the universe.



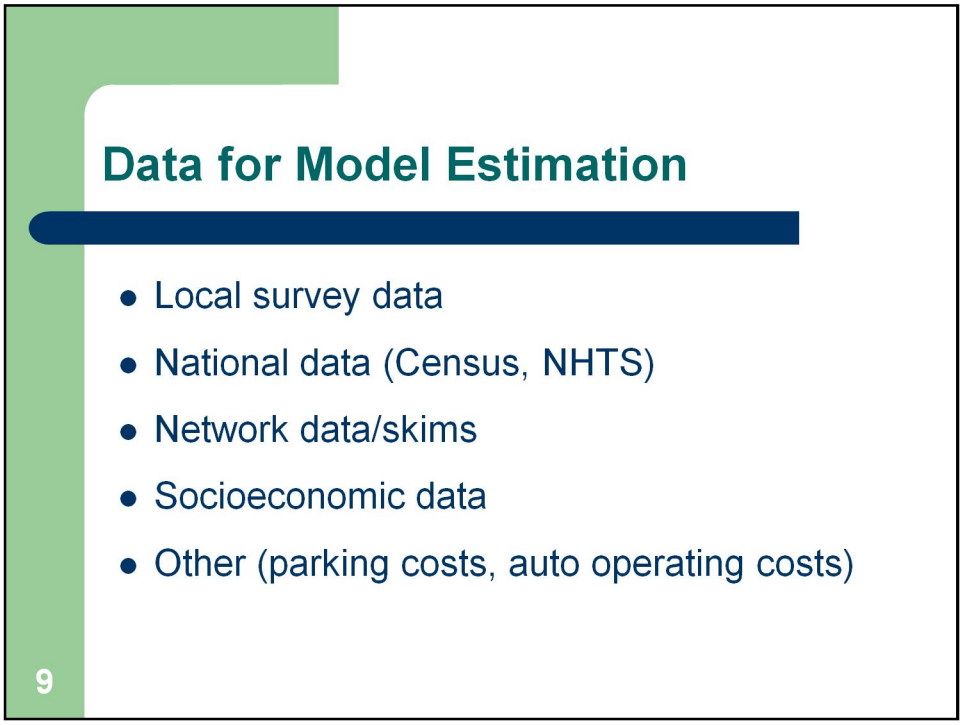
Key Message: Data Quality

Details:

Errors are virtually unavoidable in any data source. The data errors could arise during the data collection process itself through incorrect transcription of reported data, or through incorrect imputation of correctly transcribed data.

Even if the transcription and imputation are both right, data could be outdated and may not capture current travel behavior.

Another major problem with data is that it may not be representative of the true universe. This could happen in one or more ways. First, the data sample may not be large enough, thereby providing statistically insignificant results. Second the sampling plan may not be reflective of the universe.

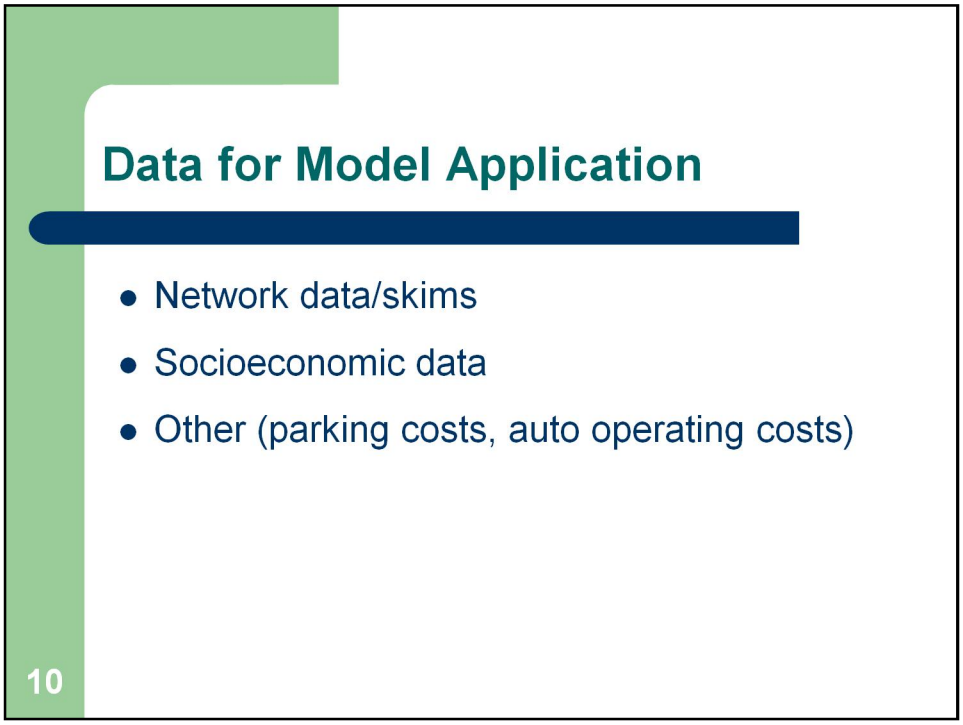


Key Message: Data Sources for Model Estimation

Details:

Some common data sources for model estimation are as follows:

1. Local survey data. These are usually the most reliable estimation data because they are specific to the region under consideration. However, they can be quite costly.
2. National data source such as NHTS and Census can be used for estimation. The national data naturally have a smaller data sample for any given region than the sample obtained from say a local survey data set. However, the national data such as NHTS and census are available free of charge.
3. The Level of service data derived from highway and transit networks is extremely important for estimating trip distribution, mode choice and assignment parameters.
4. Socio-economic data are essential for the trip generation, distribution and mode choice model estimation. These data include employment, population and household cross-classification information.
5. Other information such as the zonal average parking costs, auto operating costs per mile are extremely important, especially for mode choice model estimation purposes.

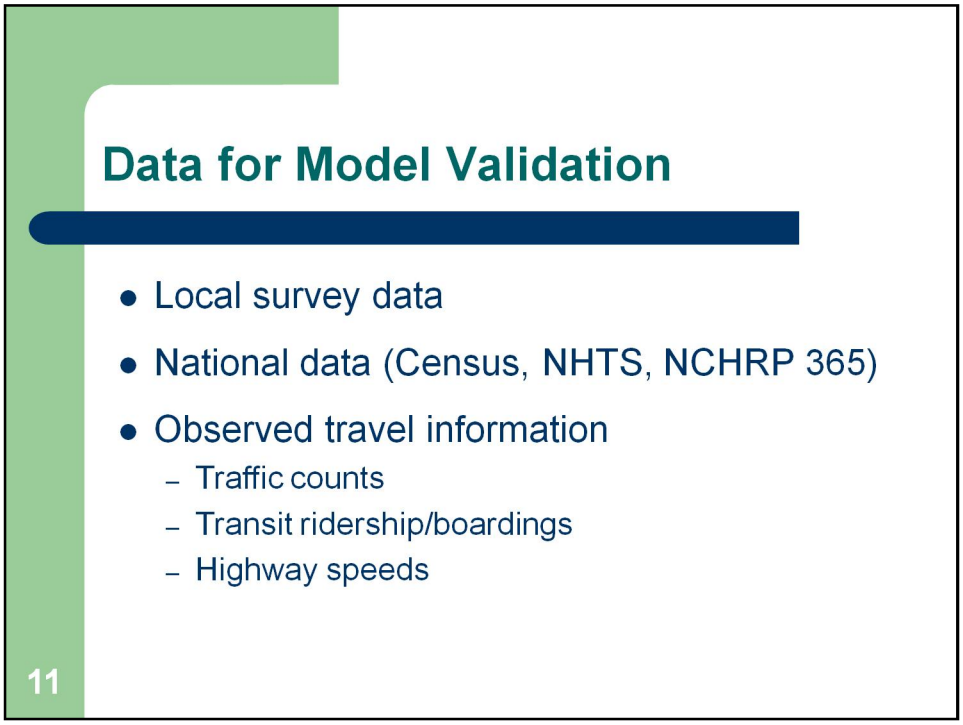


Key Message: Data Sources for Model Application

Details:

Some common data sources for model application are as follows:

1. Network data/skims are important for estimating trip distribution, mode choice and assignment parameters. In addition, they are extremely important for model application. Once the trip tables are created, they are assigned to the highway and transit networks. The modeled volumes are then compared to observed highway and transit counts.
2. Socioeconomic data are also used for application of the model. For example, zonal employment data are multiplied with the attraction rates derived during the estimation phase to generate trip attractions for each zone.
3. Parking costs and auto operating costs are commonly used for applying the mode choice models estimated during the estimation phase.

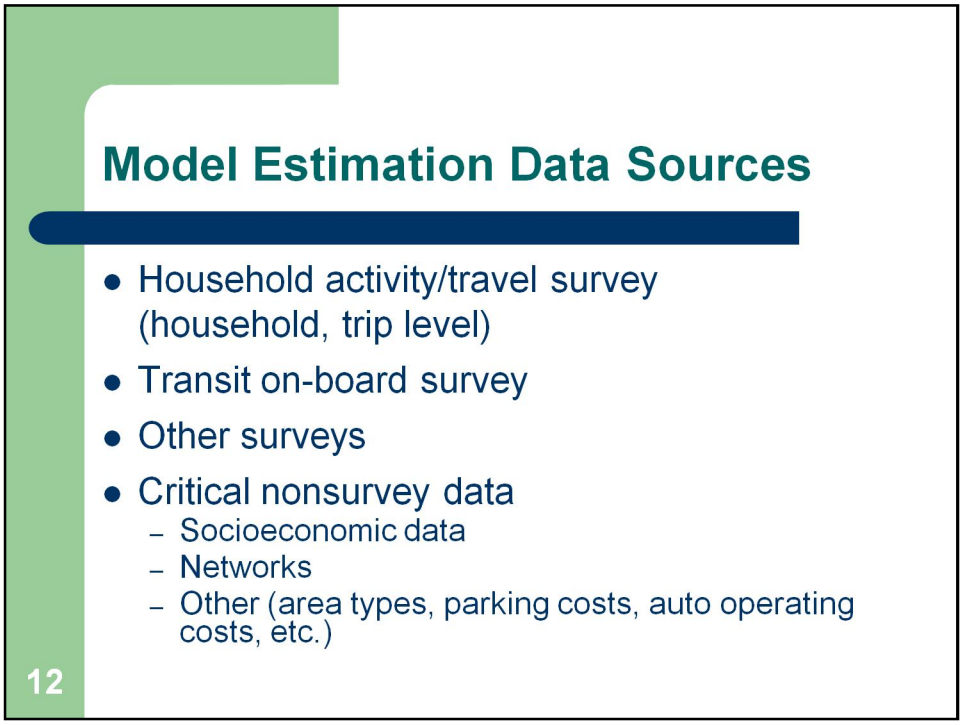


Key Message: Data Sources for Model Validation

Details:

Some common data sources for model validation are as follows:

1. As already indicated, the local survey data can be used to derive validation targets for trip generation, distribution and mode choice.
2. National data sources such as NHTS, Census can also be used as reality checks for the model patterns obtained from local models. In addition existing documents such as the NCHRP 365 can be used to compare the locally calibrated parameters to historical national numbers.
3. Observed traffic counts, transit ridership counts and highway speeds serve as validation checks for the model's trip assignment results.



Key Message: Model Estimation Data Sources

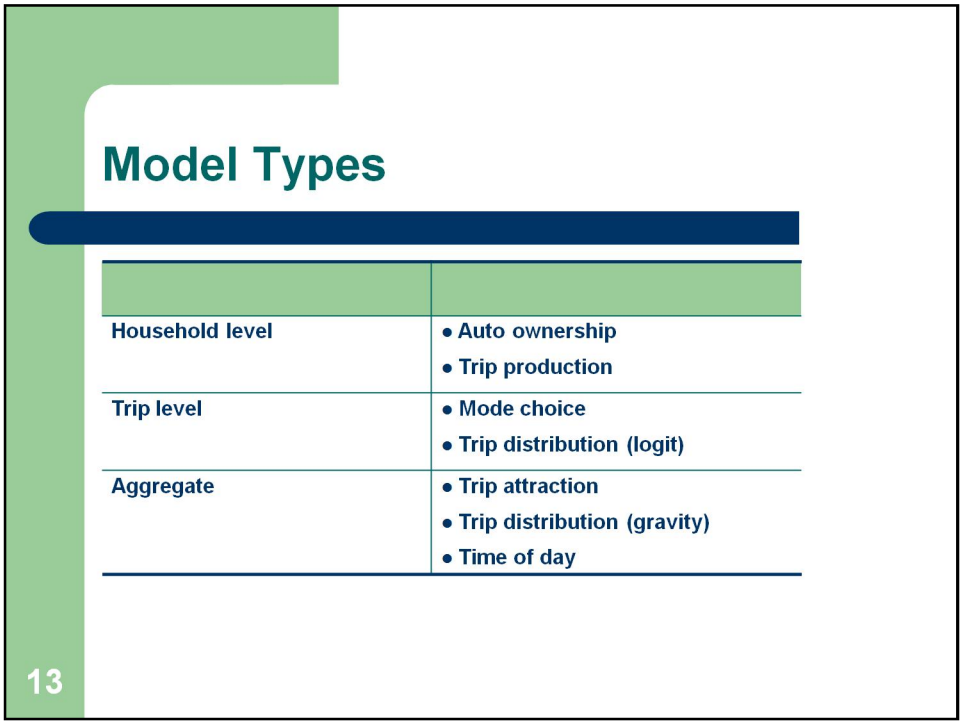
Details:

For models that are estimated from local data, the main data source is the household activity/travel survey. The data can be used to estimate parameters for trip generation, trip distribution, mode choice, time of day, and vehicle availability models, among others.

Transit on-board survey data, while also the most important data source for information on current transit ridership and validation of transit related model components, may also provide information for mode choice model estimation, since transit trip records may be sparse in household survey data.

Socioeconomic data are needed for models with variables involving demographic characteristics or densities. Network skim data are needed for models with transportation level of service variables (such as trip distribution and mode choice). Data on any other variables that will appear in the models are also needed.

The importance of data quality cannot be overemphasized! Poor data result in poor models.

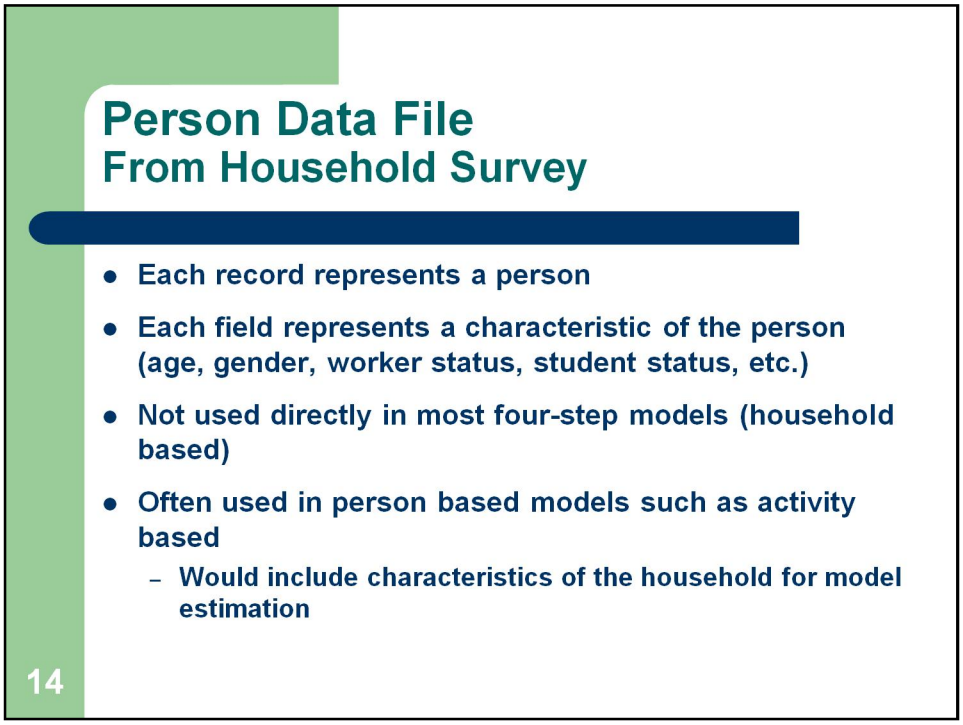


Key Message: Model Estimation Data Sources

Details:

Within a region's travel model, there are various components. These components can be thought of as representing various decisions in the daily travel process. Broadly, there are three model types depending on the decision making unit.

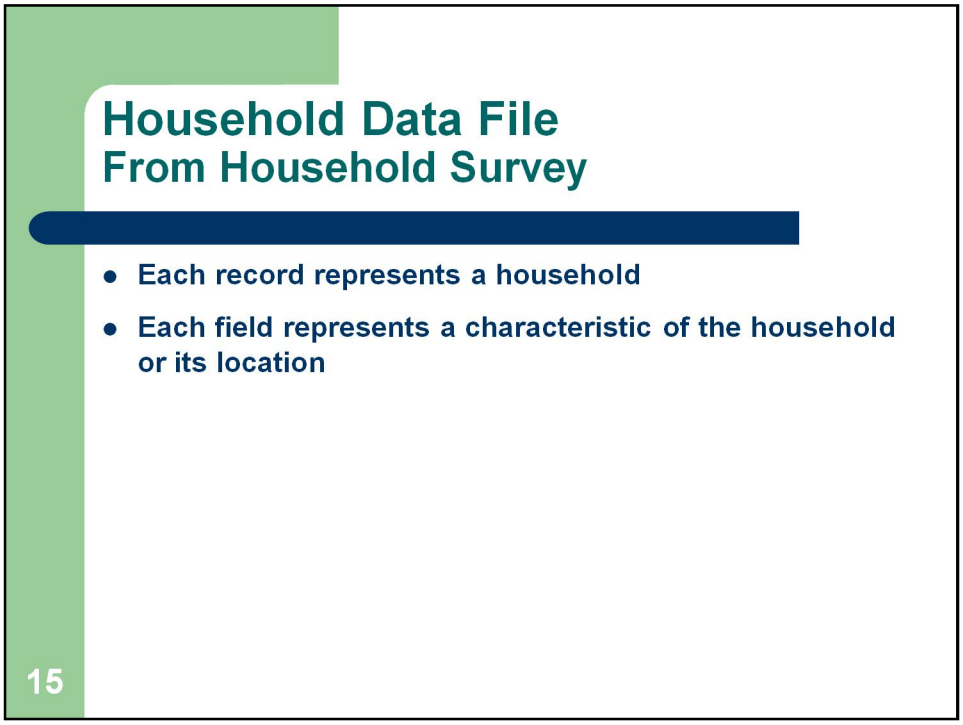
1. Household level models include auto-ownership models, which model the probability of owning various numbers of automobiles.
2. Trip level models involve decisions that are specific to each trip. For example mode choice and trip distribution models determine how to make a trip and where to go for a trip, respectively.
3. Aggregate models are typically estimated either at the zone level or sometimes at the entire region level. Trip attraction models, for example, are estimated by aggregating trips attracted to major superdistricts, and regressing these attraction totals to the appropriate employment or population measures.



Key Message: Person Data Files – What they contain.

Details:

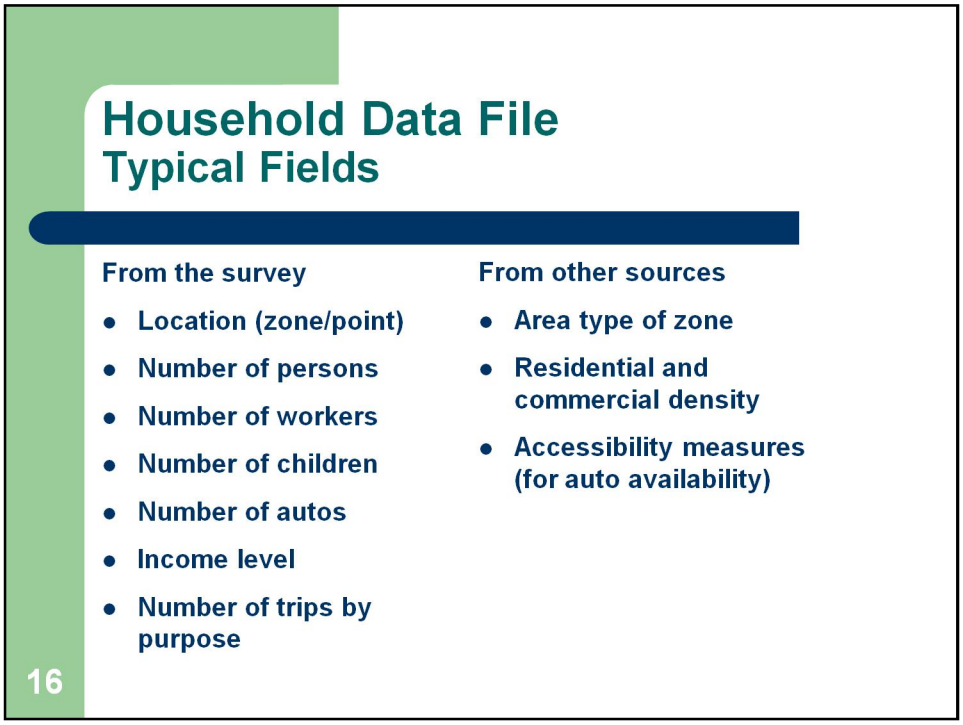
As the name suggests person files from a household survey include key socio-economic characteristics of the trip maker. Each record in these files represents a single person. Key socio-economic characteristics include age, gender, work status, student status, driver's license status and so on. These data files are typically used for estimating disaggregate models like mode choice. For the advanced model systems, such as the activity-based model systems, person files are used for estimating virtually all model components because all the components are disaggregate.



Key Message: Household Data Files – What they contain.

Details:

As the name suggests household files from a household survey include key socio-economic characteristics of the household of the trip maker. Each record in these files represents a single household. Key socio-economic characteristics include household size, number of workers, number of adults, number of vehicles, household origin zone and so forth. Household data files are one of the most commonly used data files from a household survey. They are used for both aggregate and disaggregate model estimation.

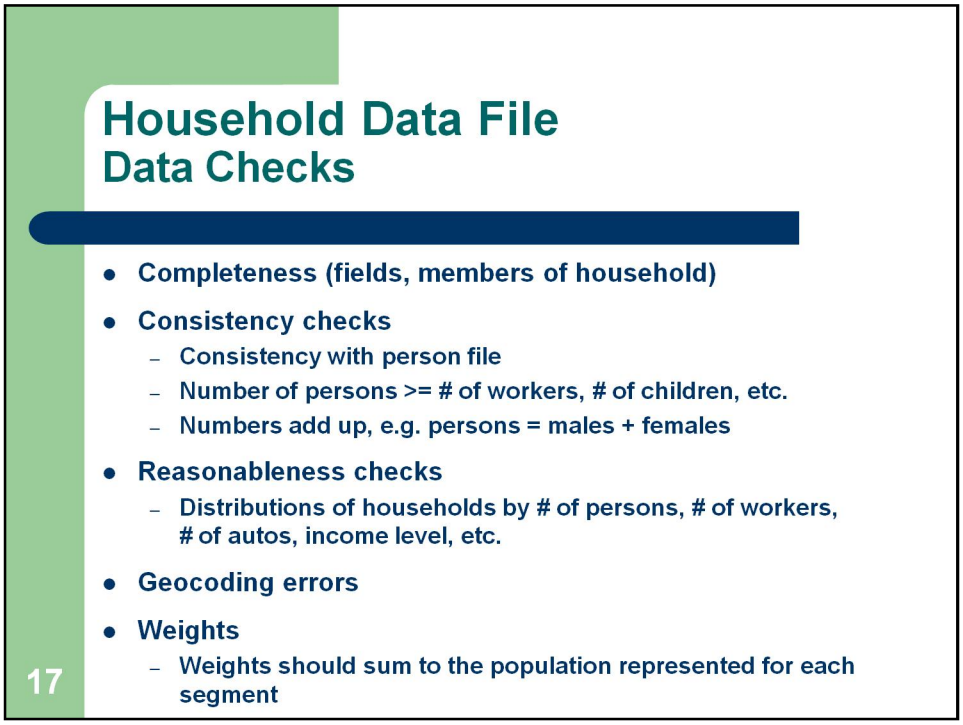


Key Message: Household Data Files – What they contain.

Details:

As the name suggests household files from a household survey include key socio-economic characteristics of the household of the trip maker. Each record in these files represents a single household. Key socio-economic characteristics include household size, number of workers, number of adults, number of vehicles, household origin zone and so forth. Household data files are one of the most commonly used data files from a household survey. They are used for both aggregate and disaggregate model estimation.

In addition to the data already present in the household data file, once can impute the area type of the zone from local knowledge and also the residential and commercial density in the zone of residence. Further accessibility measures such as number of jobs within say 15 minutes of drive time from the household, can also be imputed and attached to the household file.

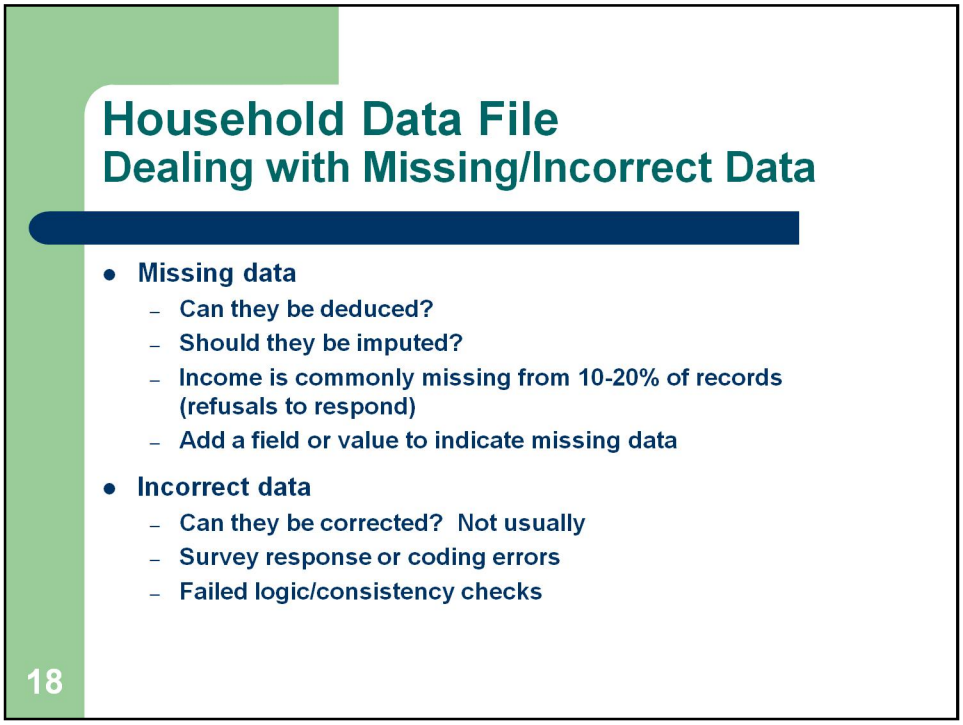


Key Message: Household Data Files – Data Checks

Details:

Before the data sets are used for estimation, it is extremely important to check them for consistency. Some common consistency checks are as flows:

1. Make sure that the household size as shown in the HH files is the same as the household size implied by the person file (at least for all those records where all the household members responded).
2. Make sure that the total number of persons as indicated by the household file is greater than or equal to the number of workers in the household, the number of adults in the household etc.
3. Make sure that the distribution of households based on household size makes sense. The average household size in the United States is a little over 2. If you notice that the average household size is 3 or 4, then there is a need to check the data for sampling biases. Similar distributions can also be made for households by vehicle ownership, income level etc.
4. Make sure that the home zone geocoding makes sense. That is ensure that the latitude and longitude information gels with the geography of the study area.
5. Make sure that the weights sum up to the total households in the region, as suggested by some other independent source such as the Census data.



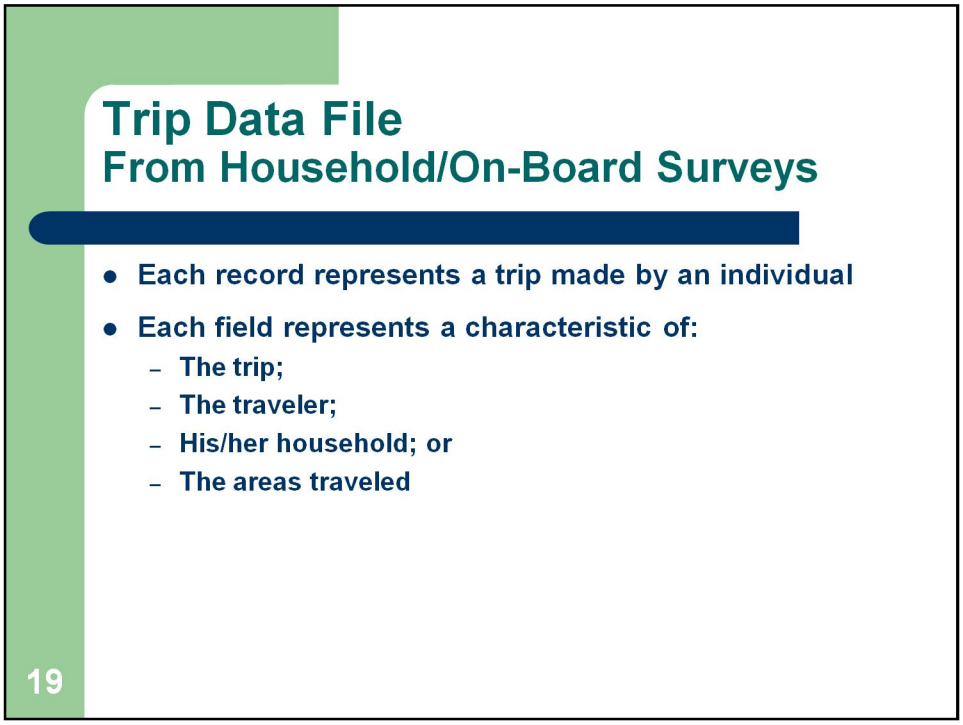
Key Message: Household Data Files – Dealing with Missing Data

Details:

Missing data are unavoidable in any survey. The critical question that we would like to ask is whether some of these missing data can be salvaged either through deduction or imputation. Income data, for example, have 10-20% missing data because some people refuse to provide any information on their incomes.

Some missing data can be deduced. For example, if the number of children in a household is missing, but the household size and the number of adults are both available, then the number of children can easily be deduced. Other missing data such as income can be imputed. Research work in the field of econometrics resulted in advanced methods that can be used to impute income data.

While missing data can be imputed, incorrect data is hard to deal with. Logic checks can indicate how bad the problem is. Ultimately this can be used to determine whether or not to use the data record at all.

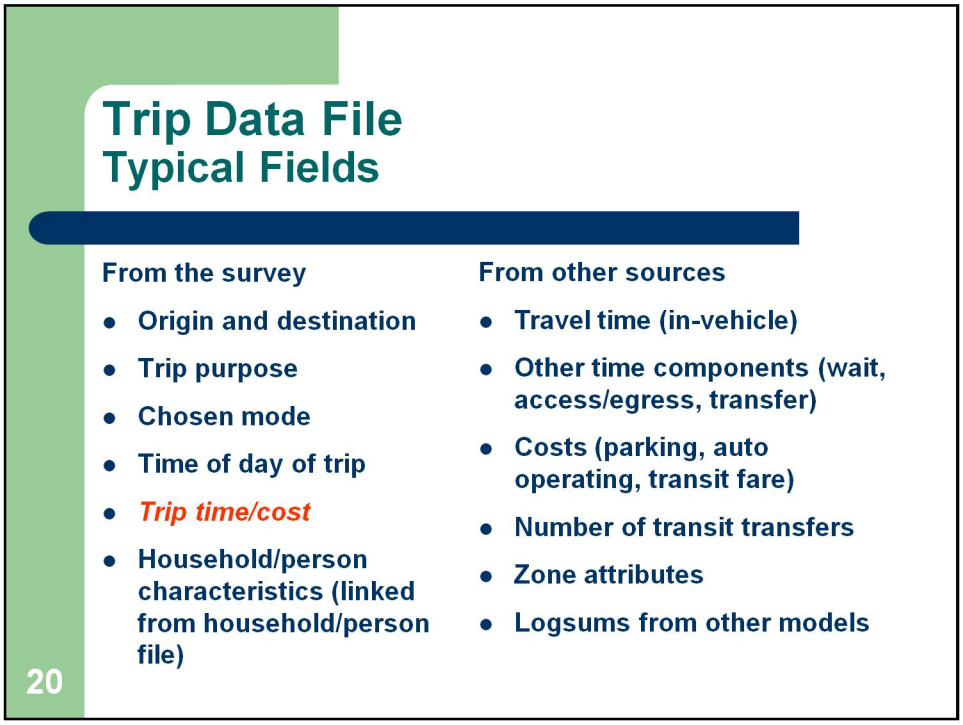


Key Message: Trip Data Files – What they contain

Details:

As the name suggests, each record in the trip file represents a single trip made by an individual. Naturally, some of the characteristics of the individual trip maker and the household are included in the trip file.

The trip file is central to all model estimation, much more so for the four-step models. A detailed list of typical fields in a trip file is shown in the next slide.

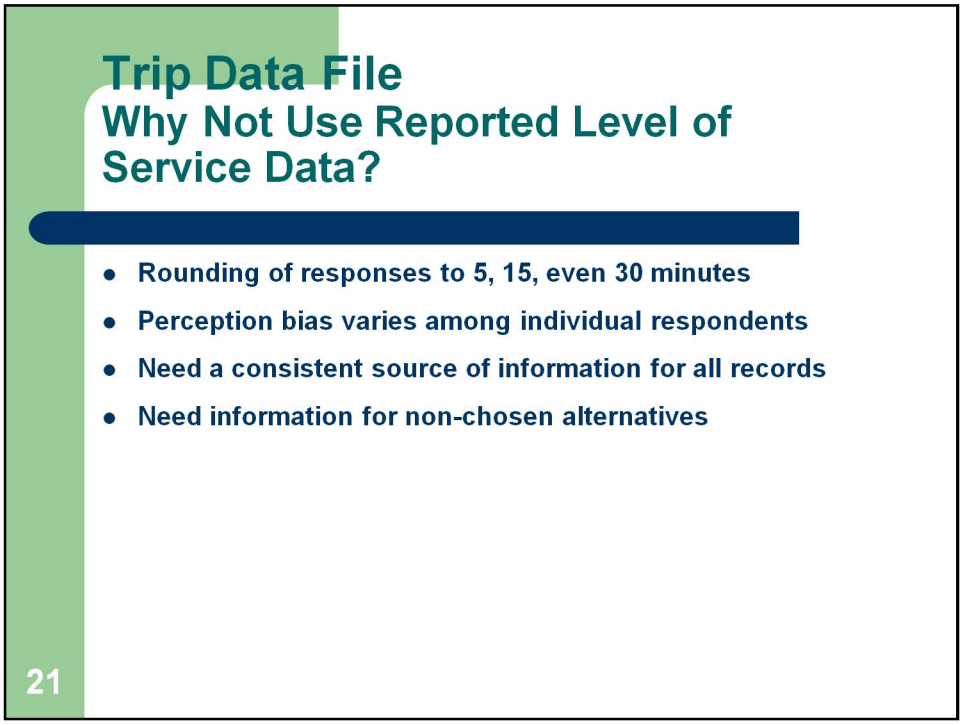


Key Message: Trip Data Files – What they contain

Details:

Some of the typical fields in a trip file are as follows:

1. Latitude and longitude information for the origin and destination of the trip. This information can be post-processed to attach TAZ information to the origin and destination.
2. Origin and destination activities, which can be used to determine the purpose of the trip itself.
3. The transportation mode chosen to make the trip.
4. The starting and ending times of each trip, which can be used to deduce the total reported time for each trip.
5. The household and person characteristics for the tripmaker.
6. In-vehicle and out-of-vehicle times can be attached to each trip record based on the origin and destination information generated in step 1. The sources of the in-vehicle and out-of-vehicle times are the highway and transit networks.
7. Zonal land use data for the origin and destination zones can also be attached to the trip files.
8. Finally, for model systems that use mode choice logsums, the logsums from the origin of the trip to each destination can also be attached to the trip record.

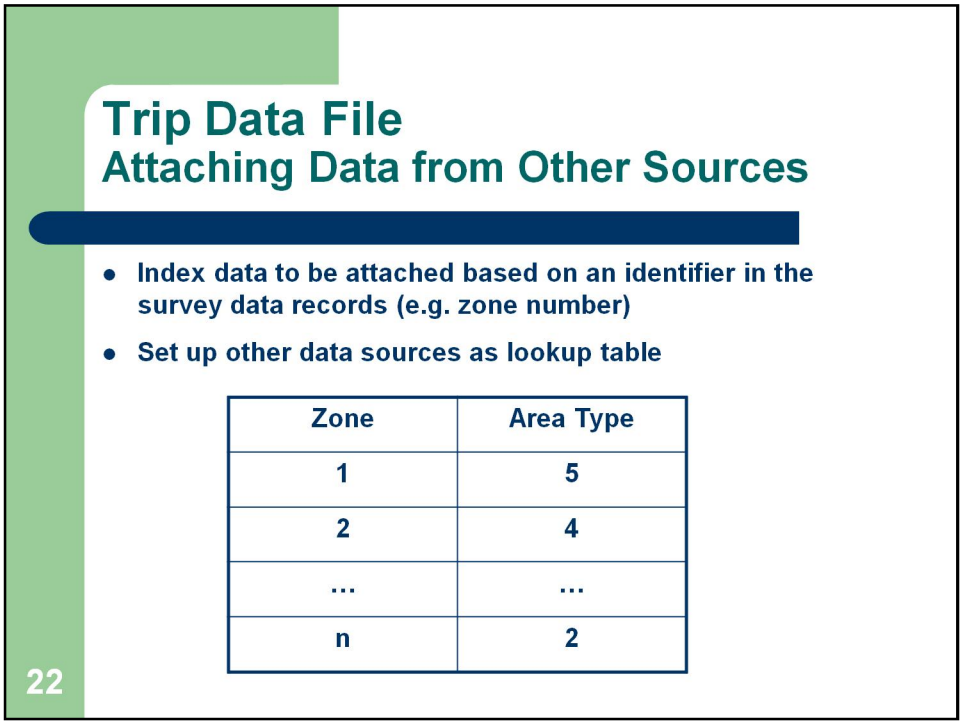


Key Message: Trip Data Files – Some Issues with Reported Data

Details:

As we just saw, many trip files include information on the start and end times of the trip. A common question that comes up while preparing data sets for estimation is whether the reported level of service data can be used for model estimation. Reported level of service data have several problems.

First, many respondents tend to round off their reported times to the nearest 5,15 or even 30 minutes. This can lead to inaccuracies in estimating total trip times. Second, respondents may perceive their travel time as more or less than their actual travel time. Therefore, the reported times may incorporate these biases. Third, the reported level of service information is obtained from each trip maker individually. In other words, there are as many sources as there are respondents. To bring all of the level of service data to a common platform, we need a consistent source of information, such as the network skims. Finally, the respondents report travel times only for the actual mode of transport that they used. For model estimation, we also need information on the other modes available but not used by the respondent.

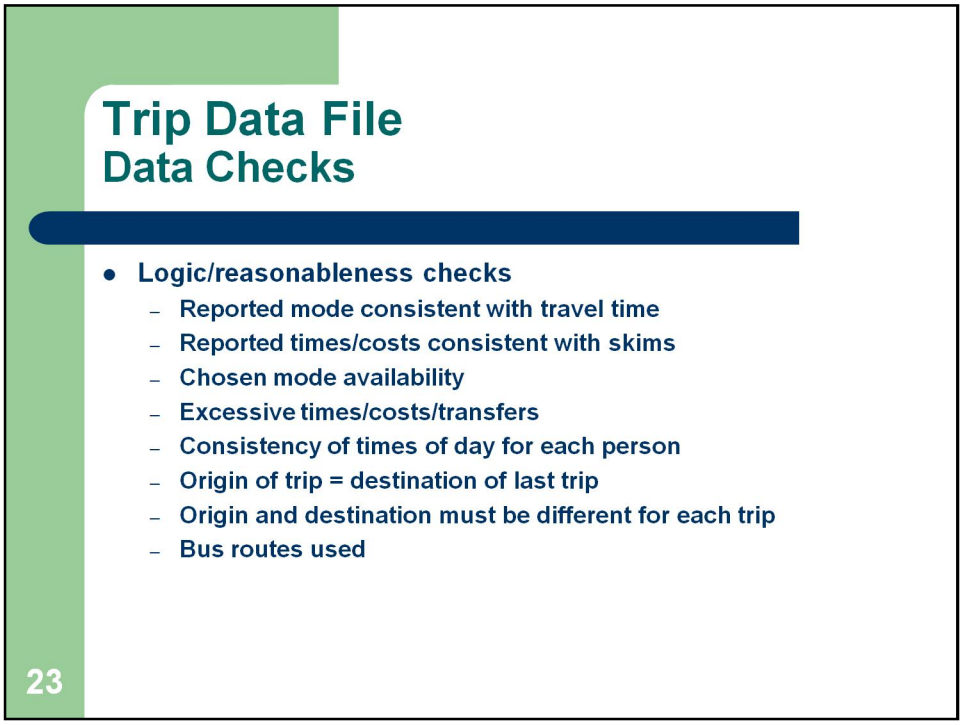


Key Message: Trip Data Files – Attaching Data from Other Sources

Details:

As we already discussed, in addition to the data available from the raw trip data files, other data can also be attached. To do this, first identify what data needs to be attached. Let's say we want to attach area type of the destination zone. Then the following steps need to be taken:

1. First, sort the trip file by the destination zone
2. Second, prepare a data file that has the destination zone and the corresponding area type.
3. Using the table in 2 as a lookup, attach the area type information to the trip file.

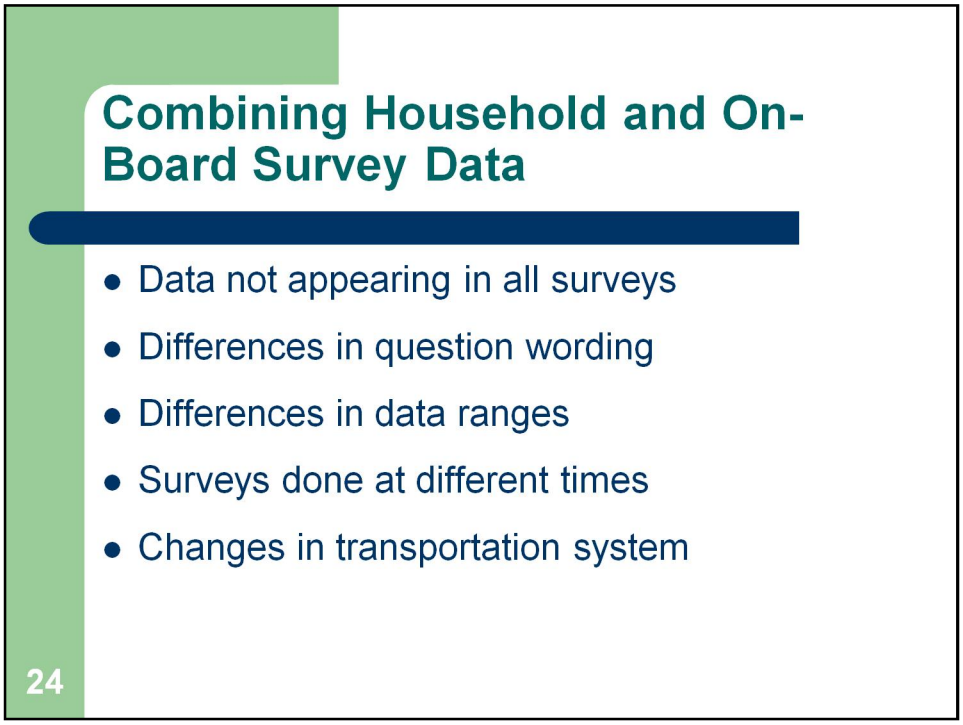


Key Message: Trip Data Files – Data Checks

Details:

Several consistency checks can be made on the trip file even before the model estimation is conducted.

1. Is the reported mode consistent with travel time?
2. Are the reported times/costs consistent with skims from independent data sources like the network skims.
3. Is the respondent's reported mode actually feasible? This is especially important for transit trips because sometimes transit may not be a feasible option for a given pair of origin and destination. Yet, the reported mode may be transit indicating that either the reported mode or the reported origin and destination information is incorrect.
4. Are the reported times/costs/transfers excessive?
5. Are the trip start and end times consistent> That is, does the trip end after it began?
6. In case of a series of trips made by the individual in a single day, does the origin of a given trip match with the destination of the previous trip?
7. Are the bus routes reported by the individual consistent with the origin and destination information?



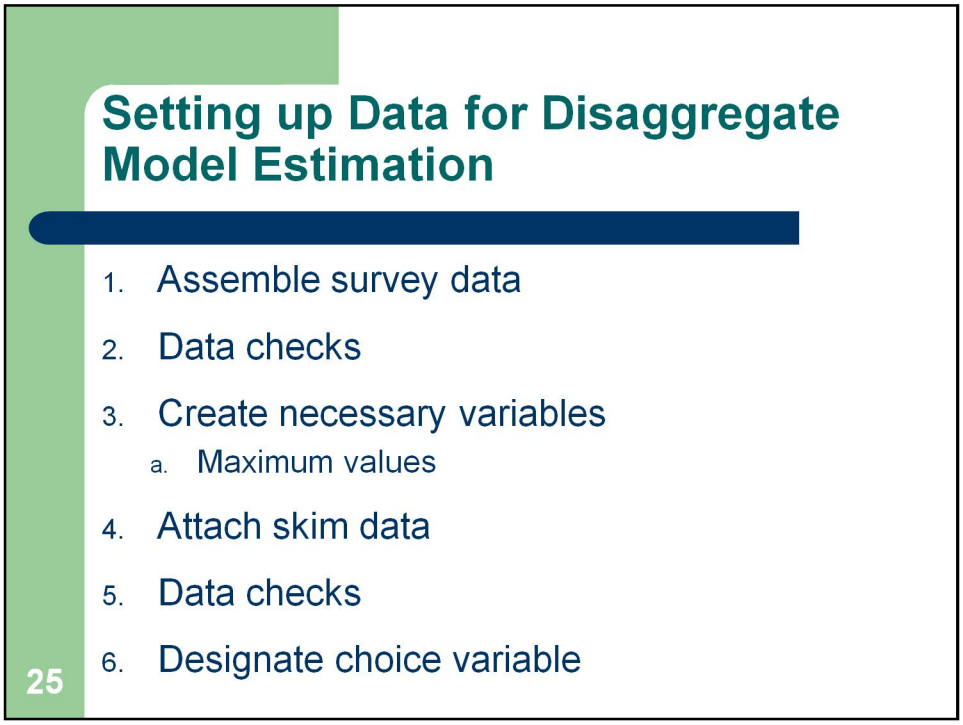
Key Message: Trip Data Files – Precautions while Combining Household and Onboard Data for Model Estimation.

Details:

For many model estimation data sets, the household survey data are combined with the onboard survey data. This is done to enrich the household data, which do not contain enough detail on transit trips, with an exclusive sample of transit trips. A few important things must be considered before combining data sets in this fashion:

Are the common variables featuring in both the household and transit onboard data consistently coded. For example, if the household file codes a gender variable as 1= Female and 2 = Male, then this same convention must be used for the transit onboard data as well. Even if the onboard data were not originally coded this way, the coding must be changed to be consistent with the household survey before merging.

Are the household and transit surveys done around the same time? If not, the costs be adjusted to represent a single yea, and of course, any changes in transportation system must also be considered.

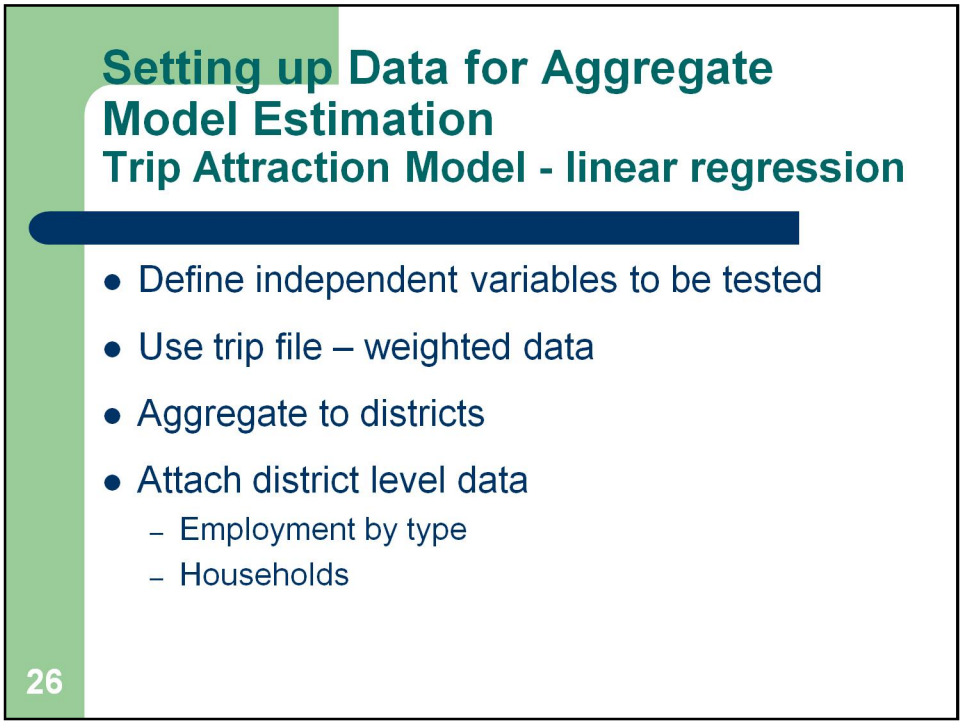


Key Message: Trip Data Files – Disaggregate Data Preparation

Details:

There are six main steps in preparing disaggregate data for estimation:

1. Assemble survey data: This entails preparing a single data file that has all the trip, household, person, zone, and level of service information necessary for model development.
2. Data checks: The consistency checks described previously need to be conducted and the incorrect and spurious records need to be flagged.
3. Create necessary variables: Any addition variables that are necessary for modeling will need to be created from the data at hand. For example, if income data are available as a single number for each record and we desire to use categorical variables instead, then we need to create the categorical variables using the continuous income variable.
4. Attach skim data: These data are typically available from the highway and transit networks as matrices.
5. Data checks: The skim data must be further subjected to data checks to see that the magnitudes of the skims conform to observed numbers.
6. Designate choice variable: Finally, a special categorical variable must be cleared to represent the choice being modeled. For example, if we are modeling mode choice, then a variable called choice, with a value of 1 = bike/walk, 2 = drive alone, 3 = shared ride, 4 = transit must be created.

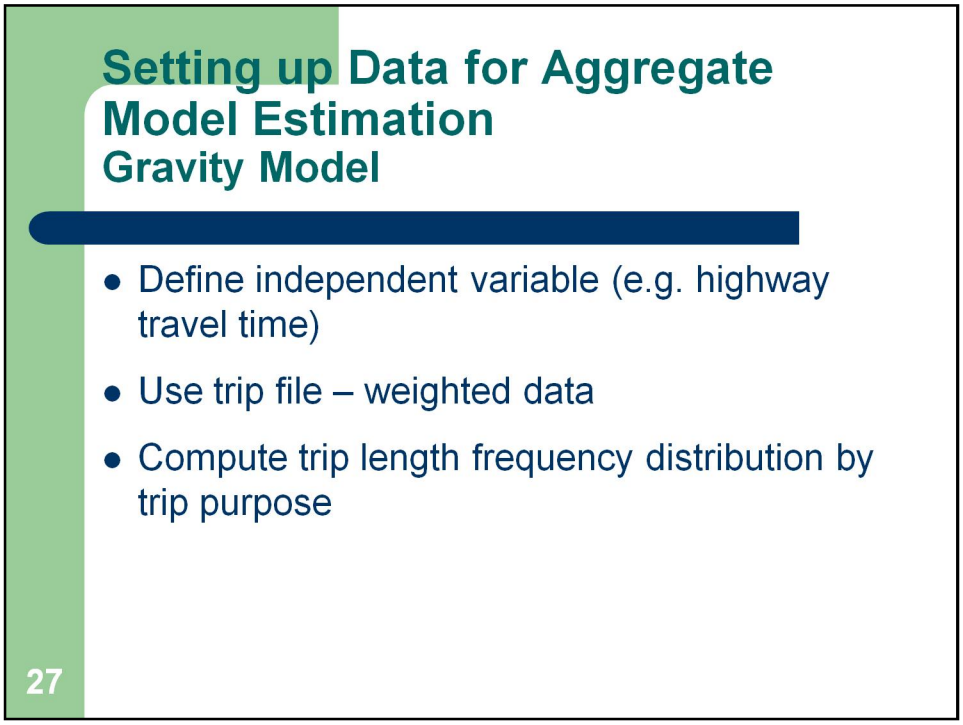


Key Message: Trip Data Files – Aggregate Model Estimation

Details:

Preparing an aggregate model estimation data file is generally simpler than preparing a data file for disaggregate models.

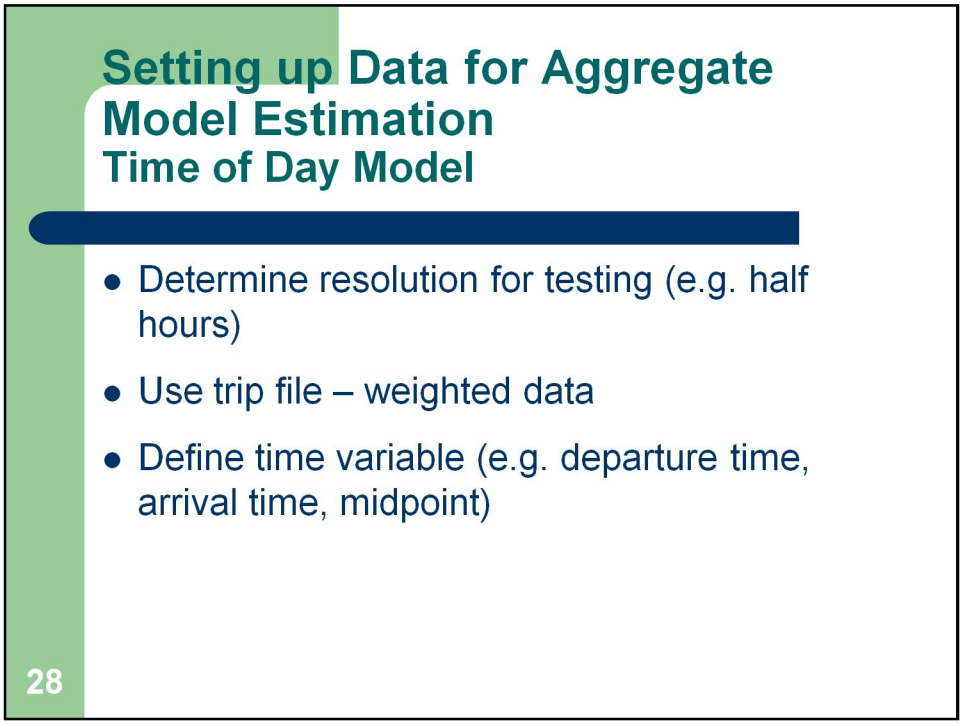
As a first step, the resolution at which the model is being estimated must be finalized. For example, for a trip attraction model, typically the linear regression models are estimated at the district level. So, the trip data must be summarized to determine the number of trips attracted to each district. Then the employment and population information must be attached to each district.



Key Message: Trip Data Files – Aggregate Model Estimation

Details:

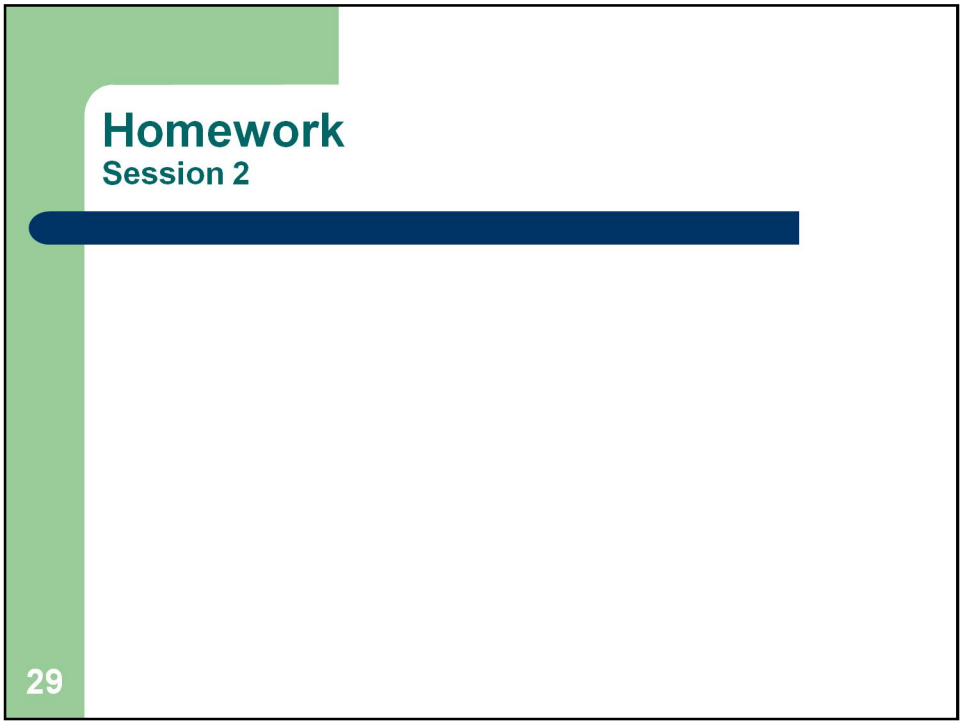
For gravity models, travel time information from the skim files must be attached to the trip file. Then, the trip file must be summarized to provide a distribution of the trip times, that is, what percentage of trips fall in each travel time bin. This information can then be used to calibrate the gravity model parameters such as the friction factors.



Key Message: Trip Data Files – Aggregate Model Estimation

Details:

Another common example of an aggregate model is the time of day factor model. Once again, the trip file forms the crux of this analysis. To develop a simple model of what percentage of daily trips occur in each time period, we need to define the size of a time period. Some models use large contiguous time periods like AM Peak, PM Peak and Off-Peak. Some other models use half-hour time periods. A second decision that needs to be made is whether the trip time period is defined by the arrival time, departure time or the mid point of the trip. Regardless of the time period definitions used, the trip start and end times must be used to summarize a frequency distribution of the trips that happen in each time period.



Key Message: Homework 2

Details:

Please refer to the homework instructions posted at the following website:

http://tmip.fhwa.dot.gov/discussions/webinars/archive/tmw/downloads/homework_2.pdf